Wyszkolenie własnego modelu AI

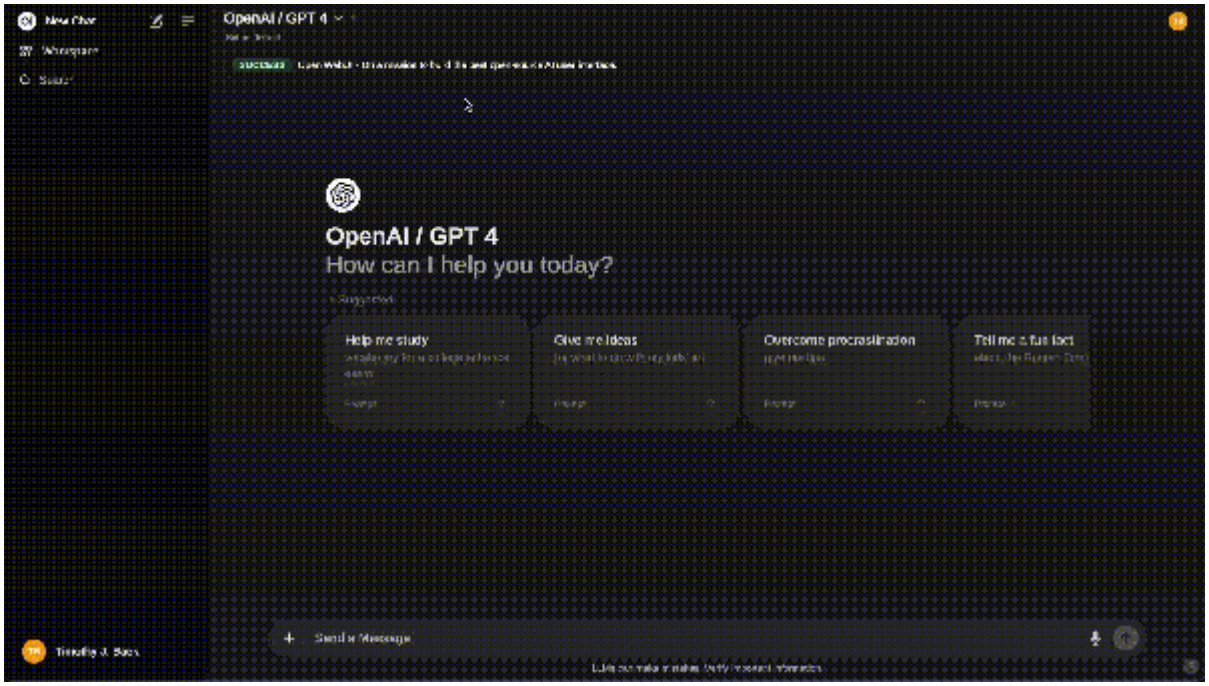
źródło: https://commons.wikimedia.org/wiki/File:Artificial-Intelligence.jpg

Wikipedia

> A large language model (LLM) is a type of machine learning model designed for natural language processing tasks such as language generation. LLMs are language models with many parameters, and are trained with self-supervised learning on a vast amount of text.

Od pewnego czasu chodzi za mną pomysł żeby wyszkolić jakichś otwarty model LLM np. Ollama, wyszkolić go wszystkimi materiałami z mojego dysku lub z tej wiki i potem opublikować go na mojej stronie jako chat bot którego można sie o wszystko zapytać.

Ollama + Openwebui



źródło: https://docs.openwebui.com/assets/images/demo-d3952c8561c4808c1d447fc061c71174.gif

Testowałem ten zestaw narzedzi i niestety ale wszystkie potrzebne pliki którymi chcielibyśmy uczyć AI trzeba wysłać przez panel webowy do modelu który trenujemy co jest strasznie mozolne. Potem model musi to wszystko przeczytać co powoduje że trwa to jeszcze dłużej. Nie jest to najlepsze rozwiązanie do takiego zastosowania jak wymieniłem we wstępie. Nie ma tutaj też możliwości podłączenia jakiegoś katalogu z plikami tak żeby AI sobie wszystko zaindeksowała a następnie odpowiadała na pytania zgodnie z tą wiedzą.

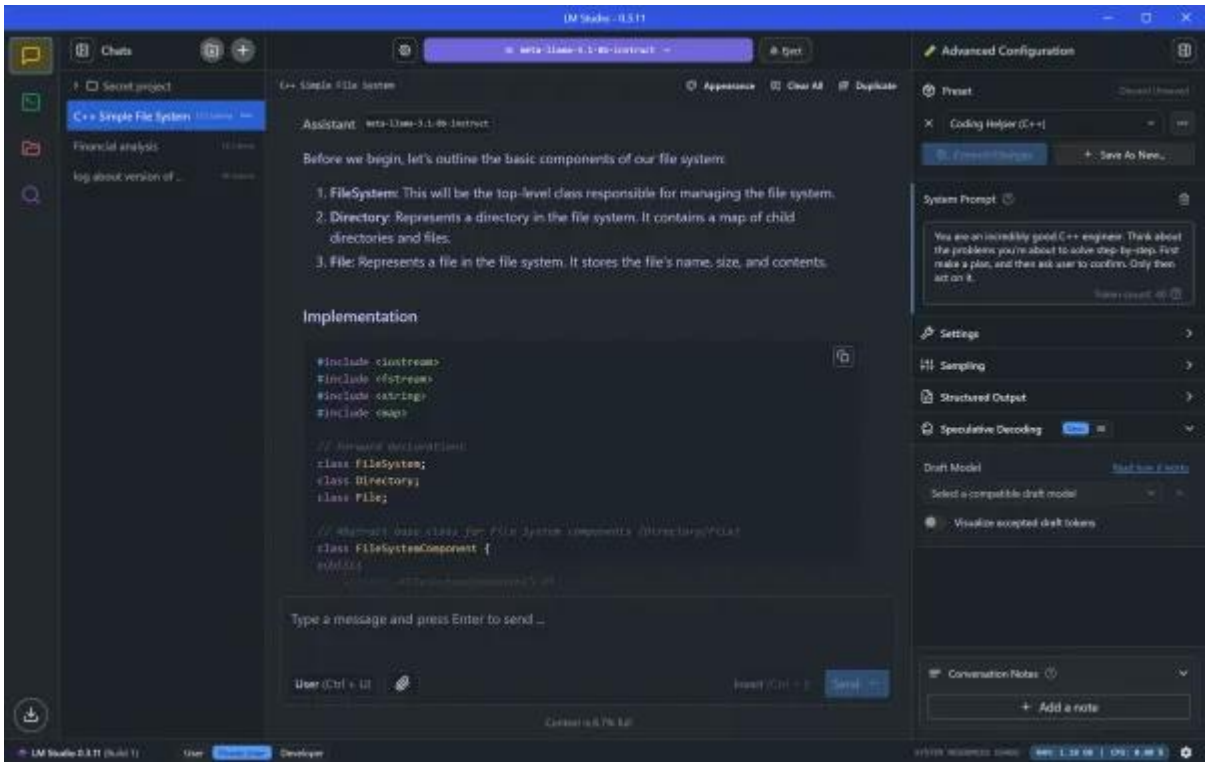
Zaleta tego rozwiązania jest jedna jest to program webowy można go otworzyć wszedzie oraz ma możliwość podłączenia różnych dostawców usług AI w jedno miejsce, co pozwala na porównywanie wyników różnych modeli AI.

Poza tym ma fajne ustawienia uprawnień do modeli oraz promptów.

Sam okazjonalnie korzystam z tego narzędzia ponieważ bardziej opłaca się płacić za poszczególne żądania do API OpenAI niż za całą subskypcje ChatGPT PLUS albo PRO.

Fajny materiał omawiający OpenWebUi na YT

LMstudio

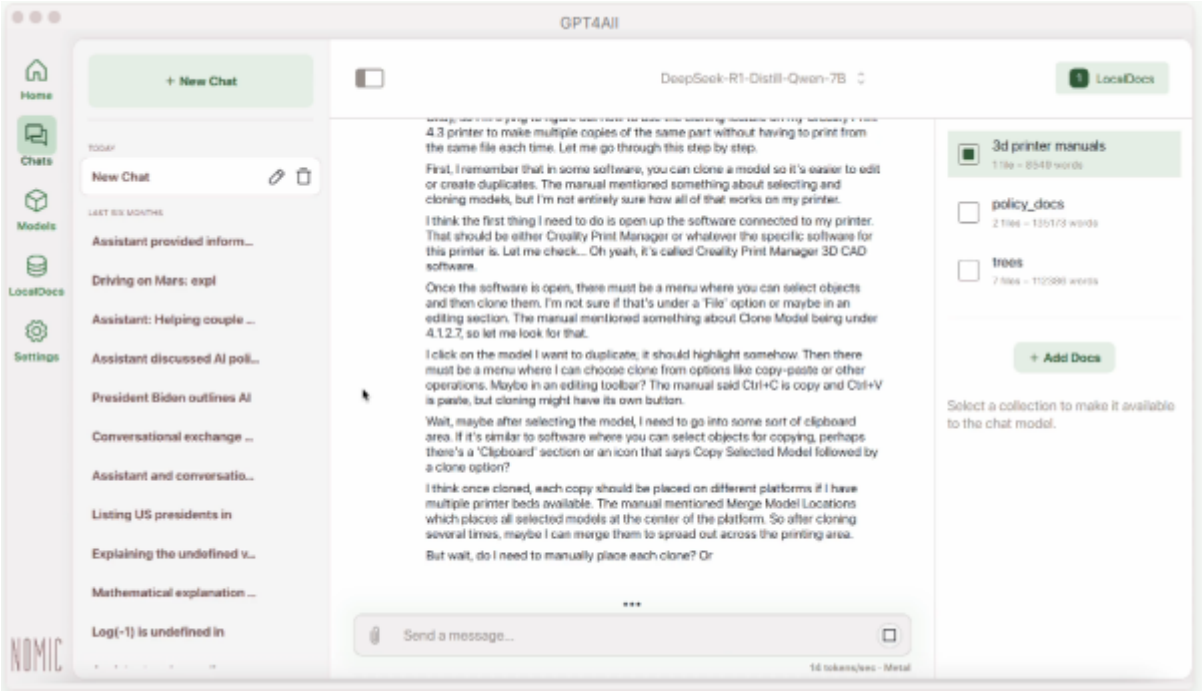


źródło:https://lmstudio.ai/_next/image?url=%2F_next%2Fstatic%2Fmedia%2Fhero-windows.2a9fa20d.webp&w=3840&q=75

To jest narzędzie które zainstalowałem na mojej głównej stacji roboczej. Jest to fajny program jeżeli chcecie na komputerze poczatować z AI na jakieś tematy poufne. Tutaj znowu jest ten sam problem jak ten z OpenWebUi, narzędzie to zachowuje się jak aplikacja Webowa przerobiona na aplikację desktop za pomocą pakietu Elektron. Ma to swoje wady, wrzucanie plików działa tak samo jak w OpenWebUi, powoduje to ten sam problem czyli nie możemy wysłać 100 plików bo cała aplikacja się zawiesi, nie ma również mozliwości podpięcia katalogu tak żeby AI go zindeksowała.

Jest to fajny program jak chcecie w łatwy sposób poeksperymentować z modelami LLM na swoim własnym komputerze.

GPT4ALL



źródło: https://www.nomic.ai/gpt4all

Od razu disclaimer nie testowałem jeszcze tego narzędzia. Natomiast ze strony możemy przeczytać że:

> Chat with Your Files Privately: Introducing LocalDocs Grant your local LLM access to your private, sensitive documents with LocalDocs. Your documents stay secure and private. Your local LLM can access your documents without an internet connection.

Czyli być może będzie to właśnie to o co mi chodzi.

W dokumentacji na stronie możemy przeczytać:

> How It Works. A LocalDocs collection uses Nomic AI's free and fast on-device embedding models to index your folder into text snippets that each get an embedding vector. These vectors allow us to find snippets from your files that are semantically similar to the questions and prompts you enter in your chats. We then include those semantically similar snippets in the prompt to the LLM. To try the embedding models yourself, we recommend using the Nomic Python SDK

Zapowiada się dobrze, z dokumentacji wynika że jest specjalne narzędzie od Nomic AI które indeksuje foldery pliki zostają zamienione na snipetty i każdy znich ma wektor zagnieżdzania, nastepnie te wektory pozwalają AI na znajdowanie snipettów z plików które są semantycznie podobne do pytań i promptów jakie podajemy modelowi. I potem te semantycznie podobne snipetty są dodawane do odpowiedzi.

Z tego co mogliśmy przeczytać jest to coś co pozwala najbliżej się zbliżyć do tego co chciałem osiągnąć.

W jednym z następnych postów będziemy robić testy tego rozwiązania i będziemy patrzeć jak sobie radzi z większą ilością danych oraz czy w ogóle się do czegoś nadaje.

— *Kacper Ostrowski 2025/05/09 11:08*